融合多源数据的科研人员画像构建方法研究*

■ 范晓玉 窦永香 赵捧未 周潇

西安电子科技大学经济与管理学院 西安 710071

摘要：[**目的/意义**]大数据时代需要将“人”数据化，科研人员也需要数据化。科研人员画像的建立，对于科研管理层全面了解科研人员的信息、客观评价其研究水平等有重要作用，可以作为分析科研人员研究行为或专家推荐的基础，提高科研管理效率。[**方法/过程**]首先提出科研人员画像的概念，认为其是描述科研人员信息的标签的集合。其次，以个人主页、知网、基金网等多个异构数据源的数据为基础，提出融合多源数据的科研人员画像构建方法，分别从科研人员的基础属性、科研偏好和科研关系三方面形式化描述了科研人员信息，并提取各个维度的标签，以可视化的方式展示其画像。最后，分别以国内外两位科研人员为例，说明了科研人员画像构建方法的可行性。[**结果/结论**]科研人员画像的构建适用于国内外的科研人员，能够全面描述科研人员信息并直观展示出来。

关键词：科研人员画像 多源数据 用户模型

分类号：G203

DOI:10.13266/j.issn.0252-3116.2018.15.004

科研人员是科技活动的主体，科研人员的相关信息是一类重要的知识资源，在科学研究、项目评审、成果转化、决策咨询等方面发挥着举足轻重的作用。2017 年 6 月发布的《2015 年我国科技人力资源发展状况分析》指出，截至 2015 年，我国科技人力资源数量持续增加，总量达到 7 915 万人，我国继续保持世界科技人力资源第一大国的地位。由于科研人员的数量不断增长，相应的问题也随之而来。一方面，虽然目前我国建立了多个科研人员信息库，但是各类信息由各个部门分别存储，数据分散，缺乏对各类信息的整合与关联，对信息的利用程度不高，科研管理部门想要了解一个科研人员，需要通过各个数据库检索，无法对科研人员有一个直观、快速的了解；另一方面，在科研评价中“唯论文”指标或过分强调论文指标的现象十分普遍[1]，不同类型的科研人员评价标准也如出一辙，忽略了科研人员其他方面的贡献。大数据时代需要将“人”数据化，科研人员也需要数据化。互联网和信息技术的迅猛发展，使得科研人员的互动和交流愈发便捷，同时科研人员的相关信息亦呈现出动态性、海量性和多源异构性等特征，对科研人员信息的有效收集、管理与分析，将有助于掌握科研现状，挖掘出科研项目在开展过程中的关键影响因素，构建出现代科技创新的“倍增器”和科学决策的“智囊团”，进而从根本上改变传统科技工作的管理与决策模式[2]。因此，本文提出科研人员画像的概念，将其用于科技人员信息大数据的分析，能对科研人员的个体特征与偏好进行充分的揭示，准确地刻画出“千人千面”，以便准确地提供个性化服务与精准推荐[3]，也为决策层提供真实有效的参考依据。

1 研究现状

近年来，越来越多的学者意识到科研人员信息的重要性，如何建立统一的科研人员信息描述框架、实现专家信息的统一描述是当前亟需研究的重要课题。在语义网、关联数据等相关研究的推动下，一些学者从大规模的人员数据中抽取人物属性和关系，构建人物本体，对人员信息及人际关系进行建模，形成便于共享和语义化描述的描述机制[4]。在科技管理领域，欧洲采

* 本文系国家社会科学基金重点项目“面向决策支持的科技管理数据深度挖掘研究”（项目编号：16ATQ007）研究成果之一。

作者简介：范晓玉（ORCID：0000-0001-8075-2154），硕士研究生；窦永香，（ORCID：0000-0001-5520-9379），教授，博士，博士生导师，通讯作者，E-mail：yxdou@xidian.edu.cn；赵捧未（ORCID：0000-0003-4026-6471），教授，博士，博士生导师；周潇（ORCID：0000-0001-5289-4876），讲师，博士。

收稿日期：2018-02-02 修回日期：2018-05-06 本文起止页码：31-40 本文责任编辑：刘远颖

用 EuroCRIS 系统构建了统一的描述模型 CERIF,将科研项目、专家、成果、机构、仪器等科技资源进行一体化关联和互操作[5]。C. Moreira 等[6]采用 D-S 证据理论和 Shannon 熵从专家成果、引文网络以及基本简介中获取专家信息。L. Yang 等[7]从专家的参与项目、奖励、发表文章、专著和授予专利 5 个方面描述了其科研成果方面的信息,并构建了基于语义和知识推理的专家系统。T. H. Trong 等[8]利用 ODP(开放式分类目录搜索系统)建立学者语义搜索空间,生成用于描述科研人员的信息本体。

在国内研究中,李纲等[9]融合知识、Web 和社会网络传感器的专家特征识别方法,设计了基于多源信息融合的专家特征识别方法。王曰芬等[10]将先进的专家检索技术、社会网络分析技术和可视化技术引入到专家库的构建中,从专家的基本特征与关系方面描述了科研人员信息,并设计了科技咨询专家库的构建流程。宋培彦等[4]以知识组织理论为基础,构建科技专家语义模型,并采用 RDF 进行形式化描述和实证研究,最终生成具有较强规范性和语义关系的专家信息库。陆伟等[11]通过网络数据库、搜索引擎、专家推荐表等渠道获取专家个人及关系信息,分别完成了专家组织工作和系统构建。J. Tang 等[12]人设计的 AMiner 系统,从海量文献及互联网信息中自动获取研究者相关信息并建立研究者描述页面,提供搜索、学术评估、合作者推荐、审稿人推荐、话题趋势分析等多样化的服务。

总体而言,无论在国外还是国内,对于科研人员信息的描述已经取得了相应的成果。但是,我们发现,有的研究中专家信息的描述局限于学术资源,描述角度单一,也缺乏对科研人员研究兴趣或科研关系的深度挖掘,描述信息不全面[4,9];其次,虽然有的研究对专家信息进行了全面的描述,但是科技管理者想要了解科研人员信息,还需要从数据库中一一查看,无法对其有一个快速直观的认识[10,12],针对上述问题,本文研究并提出科研人员画像概念及其构建方法。

2 基于多源动态数据的科技人员画像构建过程

多源数据是指由不同的用户和不同的来源渠道产生、具有多种呈现形式、描述同一主题的数据[13]。科研人员的数据来源往往有多个,且以不同的形式和角度描述了科研人员,这些不同来源的信息相互补充,以全面地描述科研人员。这些数据有的是静态的,有的是动态的,其中个人信息相对稳定,属于静态的信息,而成果信息、研究兴趣等会受到周围环境或需求的影响而发生变化,因此是动态的信息。

用户画像的概念最早由 A. Cooper 提出。他认为"用户画像能够代表一个真实的用户,是利用用户的真实数据而建立的用户模型"[14]。本文认为科研人员画像是指根据科研人员的社会属性、科研习惯、科研行为等信息抽象出的一个标签化、形式化的用户模型。参照用户画像的构建过程,本文提出科研人员画像构建方法,见图 1。

科研人员画像是一个动态更新的过程,通过定期收集科研人员的各类信息达到动态更新的效果。科研人员画像构建过程如下:首先从多个数据源收集科研人员信息,包括人口属性数据、科研成果数据、科研偏好数据等,通过数据预处理生成构建画像可用的数据,将数据存储在科研人员信息库中,然后用向量的形式将科研人员信息形式化地表示出来,构建科研人员模型,最后根据模型把科研人员各个维度的信息标签化,更新科研人员的信息标签,并使用可视化工具将科研人员画像呈现出来。

2.1 数据收集

从各类数据源中获取的我国科研人员的数据主要有人口属性数据、科研成果数据、科研行为偏好数据、科研合作数据、科研社交数据。

人口属性数据是指科研人员的人口统计学特征,包括姓名、性别、出生年月、联系电话等。科研成果数据是科研人员信息不可或缺的一部分,是指科研人员在科学研究过程中产生的具有学术意义的成果,包括期刊论文、会议论文、学术专著、专利、会议报告等。科研行为偏好数据代表了用户对某一研究主题的兴趣,科研合作数据代表了与其他人合作产生的数据,二者都是通过对科研成果信息的分析得到的。科研社交数据是科研人员在学术社交网络中产生的数据。各类数据具体的内容以及收集方式见表 1。

2.2 数据预处理

从各个来源获取的数据不能直接用来构建画像,需要对这些原始数据进行预处理。表 2 列出了各个数据类型的原始数据的特点及存在的问题。通过预处理使之转化为可用于构建科研人员画像的数据。

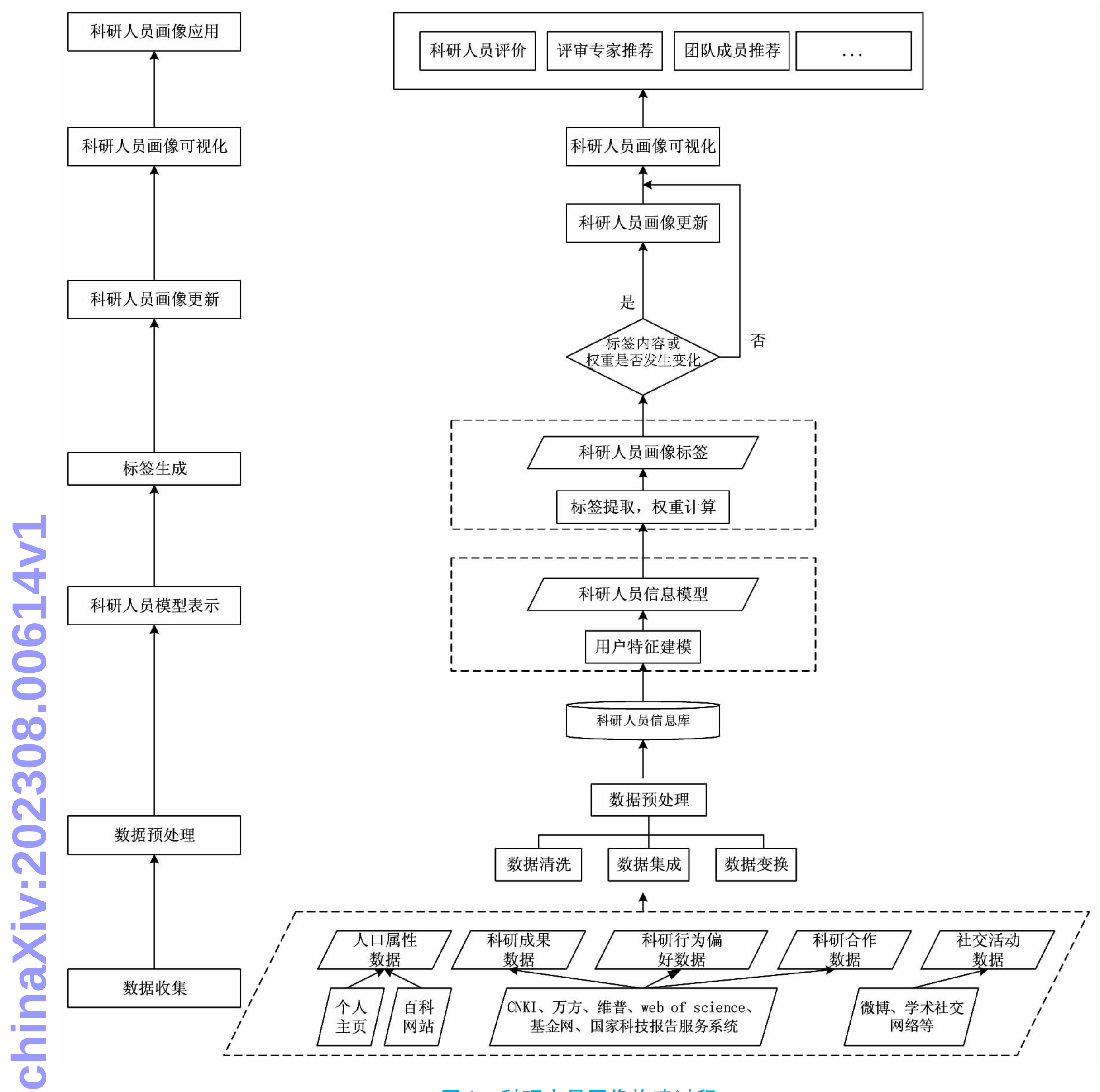


图1 科研人员画像构建过程

表1 科研人员信息来源

数据类型	数据来源	主要信息	收集方式
人口属性数据	百度百科、个人主页	姓名、性别、出生年月、联系电话、邮箱、通讯地址、工作单位、职位、职称、学历及对应的专业、工作经历	人工收集 爬虫软件
科研成果数据 科研行为偏好数据 科研合作数据	中文学术数据库、外文学术数据库	期刊论文、会议论文、学术专著、专利、会议报告、标准、软件著作权、科研奖励、人才培养、举办或参加学术会议、成果技术转移、其他重要科研成果	数据库导出 人工收集 爬虫软件
科研社交数据	学术社交网站	好友列表	爬虫软件

表2 各类数据特点及存在问题

数据类型	原始数据特点	数据存在问题
人口属性数据	文本、图片等非结构化数据	数据缺失、数据重复、非结构化、同名异义
科研成果数据 科研行为偏好数据 科研合作数据	结构化数据	数据不一致、数据重复、同名异义
科研社交数据	结构化数据	

非结构化的数据主要存在于人口属性数据中。通常科研人员的基本信息介绍会以文本的形式来描述，为了便于存储以及科研人员画像的构建，首先要将文本数据转化成结构化数据，这就需要对文本中的命名实体进行识别。命名实体主要包括名字实体（组织名、人名、地名）、时间表达式（日期、时间）和数字表达式（货币值、百分数等）。其中，针对组织名、人名和地名

的识别，由于其具有开放性和发展性的特点，识别难度比较大。目前国内有多个开源的中文语言处理工具可供直接调用实现命名实体识别，比如复旦大学研发的 fudanNLP[15] 和中国科学院的 NLPIR 分词系统[16] 都可以通过 Java 调用来实现，哈尔滨工业大学的 LTP 系统[17] 还提供了 Python 接口，可以直接用 Python 调用其封装成的 pyltp 模块实现命名实体识别，从文本中提取科研人员的信息。

结构化数据中存在有数据缺失、数据重复以及同名异义的问题。对于数据缺失问题可通过百度搜索引擎或咨询本人等途径对该信息进行补充或完善；对于数据重复问题，则需删除冗余信息，保证数据的唯一性，余下的信息相互补充；对于同名异义问题，需要通过人名消歧来解决。人名消歧目前常用的方法有基于聚类的消歧和基于实体链接的消歧，在这一方面已经有了大量的研究[18-20]，并较为成熟。

数据集成是将多个数据源中的数据整合起来统一存储到一个数据库中。在原始数据中，由于不同来源的信息采用的元数据标准不同，对科研人员信息缺乏统一的描述，同一属性在不同的数据库中会使用不同的字段名，笔者使用统一的字段对科研人员信息进行描述，具体字段如表 3 所示：

表 3　科研人员信息类别与相关字段

数据库种类	信息类别	字段
基本信息数据	个人信息	姓名、性别、出生年月、荣誉称号、研究方向
	教育信息	就读学校、就读学位、就读专业、就读时间
	工作信息	工作单位、工作时间、职位、职称、行政职务
	通讯信息	通讯地址、联系电话、Email
科研成果数据库	科研成果信息	期刊论文、会议论文、学术专著、专利、会议报告、标准、软件著作权、科研奖励、人才培养、举办或参加学术会议、成果技术转移、其他重要科研成果
科研偏好数据库	科研兴趣信息	兴趣主题、研究时间、相关成果
科研关系数据库	合著关系信息	合作者合作时间、合作成果
	社交关系信息	关注好友、关注主题

根据科研人员数据类型，笔者分别用 4 个数据库存储这些数据，分别是基本属性数据库、科研成果信息库、科研偏好数据库、科研关系数据库，各个部分之间相互关联，其中基本属性数据库存储科研人员基本的人口统计学属性数据，科研成果信息库存储各类成果信息，科研偏好数据库存储科研人员感兴趣的主题，科研关系数据库存储与科研人员产生合作关系和社交关系的科研人员信息。至此，科研人员信息已经全部转化为构建科研人员画像可用的形式，为画像的构建奠定了数据基础。

2.3　科研人员画像模型构建

目前，在对科研人员信息的描述研究中，存在以下两个问题：①数据描述单一，单一地从科研人员的学术成果方面描述科研人员，忽略其他方面的信息；②对科研人员信息没有一个直观的展示。针对这两点问题，本文融合多数据源提出了科研人员画像模型，并将该模型进行了实例化，将各个维度的标签存储在对应的标签库中。

2.3.1　科研人员画像模型　科研人员画像是一个多维度、多层次的用户模型。根据科研人员信息库中数据的类型，本文定义一个三元组作为用户信息的向量空间表示：

$$User = <Demographics, Interests, Relation>$$

其中，代表用户的基础属性维度，作为用户信息的向量空间表示。*Demographics* 代表用户的基础属性维度，*Interests* 表示用户的科研偏好维度，*Relation* 代表用户的科研关系维度。多层次的科研人员画像模型见图 2。

2.3.2　标签提取及权重计算

（1）基础模型标签提取。在 2.3.1 节中，笔者用 $Demographics = <BaseInfo, Edu, Org, Message, Achv>$ 来表示用户的基础属性模型，由人口统计学维度和科研成果维度组成。由于在科研信息库中，科研人员的基础信息表示精炼，可以直接采用数据库中的信息作为标签，科研成果信息的标签采用成果的标题来表示。

（2）科研偏好标签提取。科研偏好向量模型为：

$$Interests = \{(Topic_1, t_1), (Topic_2, t_2), (Topic_3, t_3), \cdots, (Topic_n, R_n)\}$$

其中，$Topic_n$ 表示科研人员的第 n 个兴趣主题，t_n 表示用户对第 n 个主题的兴趣度，t_n 越大，表示用户的兴趣度越高。由于关键词是对文章内容的高度凝练和概括，使用关键词可以作为研究主题的主要表征。本文将科研人员发表文献的关键词作为科研人员的 *Topic* 标签，并依次计算其权重，权重计算如下：

由于科研人员的研究主题不是一成不变，会随着周围客观环境或主观兴趣发生改变，因此，研究主题标签权重的计算综合采用其偏好权重和衰减权重来表示。偏好权重是指该标签在所有标签中占的比重，用 $\omega_i = \frac{n_i}{N}$ 表示，其中 n_i 表示标签出现的次数，N 表示标签总数。衰减权重方面，借助 Y. Cheng 等[21] 提出的自适

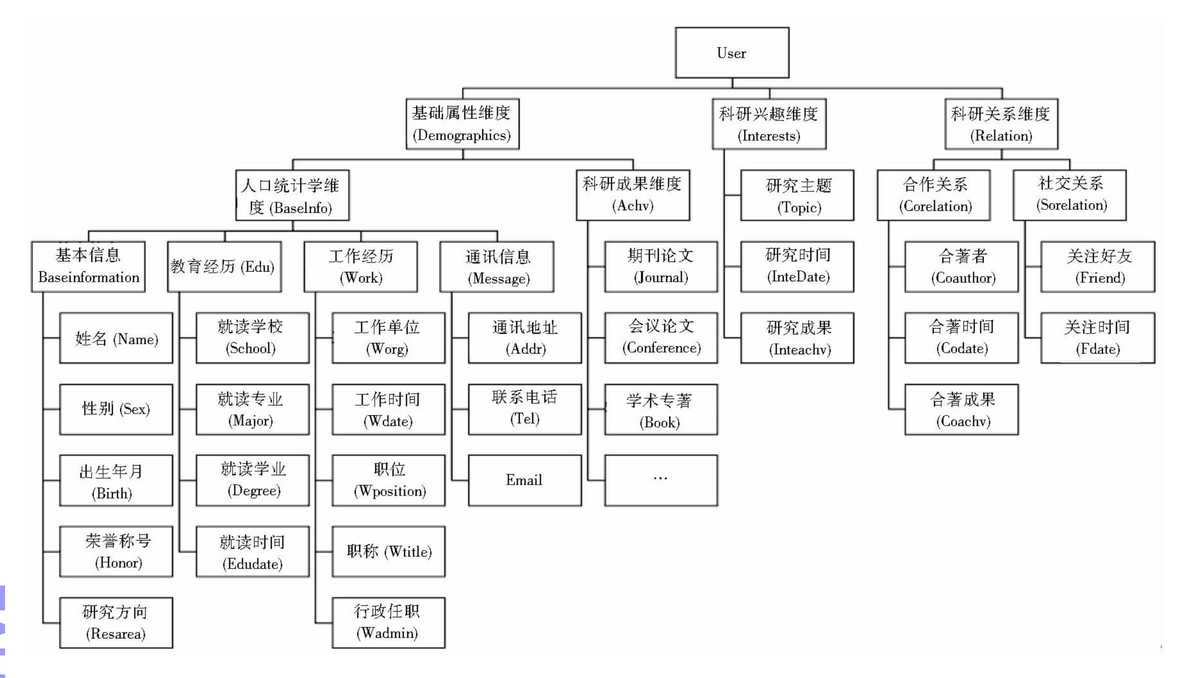


图2　多层次的科研人员画像模型

应指数衰减函数来表示科研人员对某一标签兴趣的衰减程度：

$$\theta_i = e^{-\frac{\ln 2 \times (t - est)}{hl}} \tag{1}$$

其中，t 为当前时间，est 为该标签出现的最早时间，hl 为兴趣主题衰减的半衰期。科研人员对主题的行为周期即对这一标签的研究周期越短，hl 越小，兴趣度下降越快，否则兴趣度下降越慢。因此，科研人员近期研究的主题被赋予更大的权重，时间越久远的兴趣度越小。

综合标签的偏好权重和衰减权重，某一标签的综合权重表示为：

$$t = \lambda\omega_i + (1-\lambda)\theta_i \tag{2}$$

其中，λ 是调和因子，用来调节偏好权重和衰减权重的大小，这样既可以得到用户对研究主题的偏好程度，又考虑了时间的因素，反映了科研人员兴趣的漂移。

（3）科研关系标签提取。科研关系向量模型 $Relation = <ReTag, R_{ui}, R_u>$ 中，Re*Tag* 为科研关系中的节点标签，R_{ui} 为用户 u 和用户 i 关系权重，在科研关系图中用节点间连线的粗细来表示，R_u 为用户 u 在关系中的贡献大小，在科研关系图中用节点的大小表示。

Re*Tag* 节点标签直接采用与其产生合著关系的科研人员姓名来表示，获取科研人员发表文章中的所有作者姓名，去重之后作为科研关系向量模型中的 Re*Tag*。

在计算 R_u 时，根据用户 u 在这些文章中署名位置的不同来计算其贡献值。本文参照作者贡献率等级分配法[22]，按照作者的署名顺序分别计算每位作者的贡献权重，最终将作者在每篇文章中的权重叠加得到该作者的节点权重。等级分配法是指合著文献中每位著者的权重按其在文献中的排名由先至后依次递减，假设某篇文章有5位合著者，那么第一位到第五位合著者的贡献度分别为5/15、4/15、3/15、2/15、1/15。因此，某篇文章k中合著人数为 n，排名第 i 位的作者的贡献度为 ω_i：

$$\omega_i = \frac{n-i+1}{\sum_{i=1}^{n} i} \tag{3}$$

如该作者发表的文章为 m，该作者总的贡献度大小为：

$$W = \sum_{k=1}^{m} \omega_{ki} = \sum_{k=1}^{m} \frac{n-i+1}{\sum_{i=1}^{n} i} \tag{4}$$

本文中科研人员的数据来源有多个，为了体现不同来源的重要程度，根据不同来源设定该作者贡献度的加权公式为：

$$R_u = \alpha_1 W_1 + \alpha_2 W_2 + \cdots + \alpha_p W_p \tag{5}$$

其中，α_p 表示不同来源的文献占文献总数的比重。假设科研人员数据来源有 p 个，从各个来源中获取的文献数量分别为 $x_i, x_2, \cdots, x_p$ 那么

$$\alpha_p = \frac{x_p}{x_1 + x_2 + \cdots + x_p} \tag{6}$$

R_{ui} 表示用户 u 和用户 i 产生关系的强度，用户之间关系包括其合著关系，也包括社交关系，分别用

CoR_{ui}和 SoR_{ui}来表示，且 $R_{ui}=0.5CoR_{ui}+0.5SoR_{ui}$。科研人员的合著关系用其合著的文章在总的发文数中占的比例来表示，即

$$CoR_{ui}=\frac{Paper_{ui}}{Paper_u} \quad (7)$$

其中 $Papaer_{ui}$表示用户 u 和用户 i 共同发文数，$Papaer_u$ 表示用户 u 的发文总数。科研人员的社交关系用一个布尔值来表示，如果用户 i 是用户 u 的好友，$SoR_{ui}=1$，否则用 0 来表示。

2.4 科研人员画像更新

科研人员画像的更新也就是对科研人员基本信息、科研兴趣以及科研关系标签的更新。科研人员发生变化的基本信息主要有其工作单位、联系方式、通讯地址、职称及职务等。基本信息的更新可以由科研人员信息库管理者定期向科研人员发送邮件，提醒其更新自己的基本信息。当科研人员基本信息库中的数据发生变化时，其对应的标签也发生相应的变化。

科研人员的科研兴趣标签和科研关系标签是基于科研成果信息发生变化的。据统计，在 2010 年中国的 20 种科技期刊中，3164 篇论文的平均发表周期为 11.6 个月[23]，因此应定期收集科研人员成果信息，根据图 1 的流程提取新的成果信息中的关键词和合作者，作为新的兴趣标签和科研关系标签，按照 2.3.2 节中的方法计算其权重，并与已经存在的标签进行比较，如果标签内容或权重发生了变化，就用新的数据替换原来的数据，然后再对其进行可视化；如果没有发生变化，直接对原来的标签进行可视化。

2.5 科研人员画像可视化

科研人员画像可以看作是用户信息的标签云，根据不同标签的权重，用不同的大小将科研人员的信息形象直观地呈现出来。现已经有成熟的工具用于实现标签的可视化，如 Wordle、tagCloud、Tagul[24]、Tagxedo[25]等。

3 实例验证

为了证实方法的可行性，本文分别构建了国内外科研人员的画像。国内以微电子领域的某位专家 YYT 教授为例，他在国内外学术刊物和重要学术会议上发表论文数百篇；国外科研人员以华盛顿大学计算机科学与工程学教授 P. Domingos 为例，他是国际机器学习协会的联合创始人之一，其学术水平得到国内外同行的认可。本文给出了 YYT 教授的具体构建过程，P. Domingos 画像的构建过程同 YYT 教授，最终将二者的科研人员画像可视化展示出来。

3.1 数据收集

YYT 的人口属性数据来自百度百科和个人主页，百度百科作为全球最大的中文百科全书，包含了 150 亿以上的词条，几乎涉及所有的知识领域。科研人员的个人主页也涵盖了对其自身的基本信息、科研内容的介绍。用爬虫的方式获取科研人员的人口属性数据。

科研成果数据来源包括中文学术数据库、外文学术数据库。中文学术数据库主要有 CNKI 数据库、万方数据库、维普数据库，这三类数据库在内容方面有重复现象，但是又相互补充；外文数据库主要有 Web of Science 和 EI 数据库。国家自然科学基金网、国家社会科学基金项目数据库中有科研人员近年来参与的各个项目情况；国家科技报告服务系统提供了科研活动的专题报告、进展报告、最终报告和组织管理报告。最终共采集到期刊论文数据 892 条（中文 502 条，外文 390 条），会议论文数据 120 条（中文 6 条，外文 114 条），中文专利数据 519 条，培养的硕博士论文数据 208 条。

科研社交数据来源于全球最成功的学术社交网站之一——ResearchGate。利用爬虫工具八爪鱼从该网站爬取用户的好友关系。虽然该网站有用户的基本信息，也有用户的成果信息，但是为了数据的唯一性，笔者从该网站只获取 YYT 的好友列表，最终获取到数据 40 条。

同样，收集 P. Domingos 教授的各类信息数据，人口属性数据来自他的个人主页，成果数据来自 Web of science 和 EI 数据库，包括会议论文 95 篇、期刊论文 81 篇、专著 4 篇，并从 ResearchGate 获得 59 条科研社交数据。

3.2 数据预处理

在收集到的成果数据中，有些数据是构建科研人员画像不需要的数据，因此只提取构建画像所需要的字段信息，如题目、作者、关键词、发表时间。外文数据中，如出版日期采用的是英文的格式，为了数据的统一，将所有信息中的日期格式设置为“xxxx-xx-xx”的形式。此外，还存在数据重复的现象，由于笔者获取的数据保存在数据库中，利用 Distinct 命令去除重复数据。此外，数据预处理中还存在一些关键问题，主要有命名实体的识别和人名消歧问题。

3.2.1 命名实体识别　获取的人口属性数据中，有部分是文本形式的数据，如对科研人员工作经历的介绍，本文利用哈尔滨工业大学的 LTP 命名实体处理工具，

范晓玉，窦永香，赵捧未，等.融合多源数据的科研人员画像构建方法研究[J].图书情报工作,2018,62(15):31－40.

用 Python 调用其封装成的 pyltp 模块实现命名实体识别,包括姓名、籍贯、机构、出生日期等。如果直接利用 Python 中自带的分词词典,一些较长的词语比如“西安电子科技大学”就会被分成“西安/电子/科技/大学”,这样抽取的词没有实际意义。因此,在分词时自定义分词词典,以便更好地抽取科研人员信息。

3.2.2 人名消歧处理　在获取的 YYT 的英文合作者中,有出现类似于“Zhu ZY”“Wang JY”这样的名字缩写,为了更好地区分这些名字缩写对应的真实姓名,本文采用宋文强提出的科技文献中同名区分的方法——基于分布聚类的消歧算法[18]。该算法的基本步骤如下:①将每一篇文献当作一簇,计算任意两篇文章之间的相似度,得到初始的 N＊N 矩阵 D;②查找 D 中相似度最大的两个文献记录,将它们合并到一个新簇中;③重新计算新簇与其他所有文献的相似度;④重复步骤②和③,直到最后的文献簇为给定的簇数。根据中文合作者,为每一簇匹配真实的科研人员。该方法的消歧准确率达到了 90%。本文用 Python 语言实现了该算法并用于人名消歧。表 4 统计了获取的数据中人名缩写及相关文献数量,并利用该算法进行了区分。经过消歧,在“Zhu ZY”的 94 篇文献中,区分出有 26 篇的真正作者是“朱作云”,有 37 篇的真正作者是“朱振宇”,有 13 篇的作者是“朱兆义”。在“Wang JY”的 10 篇文献中,3 篇的真正作者是王建云,2 篇的真正作者是王居勇。将区分后的作者信息分别补充到对应的属性中。剩余未区分出的作者信息采用人工的办法,根据其机构信息进行进一步区分。最后将处理好的数据存入到数据库中,以供后续标签提取及权重计算。

表 4　人名缩写数据统计

名字缩写	文献数量	相同缩写作者数
Zhu ZY	94	3
Wang JY	10	2

3.3　模型表示及标签提取

在 2.3.2 中,科研人员的基础属性维度的标签直接用科研人员信息库中的信息来表示。科研成果 *Achv* 的标签用文献的标题来表示。由于科研成果较多,把近三年内的文献标题作为 *Achv* 的标签。

科研偏好标签用成果信息中的关键词表示。在统计时,由于科研人员的文献来自于中外文数据库,在统计之前需要将中英文对照,将所有英文关键词转化成中文,人工修正之后再进行统计,把使用频次较多的关键词作为科研人员的研究主题标签,并根据公式(1)和(2)分别计算权重,计算时,设置调和因子 $\lambda=0.5$,认为科研人员的偏好权重和衰减权重在综合权重中所占比例相同。衰减权重中半衰期的取值,根据关键词半衰期的定义[26]来计算。科研人员关键词半衰期是指某年度使用过的关键词最新的一半是多长时间内创建的,计算公式是 $hl=A+(50\%-B)/C$,其中,A 为累计百分比最接近 50% 那年经历的年数,B 为累计百分比最接近 50% 的那年对应的累计百分比,C 为累计百分比第一次超过 50% 的那年的当年百分比。最终以 2017 年为起始年,累计百分比最接近 50% 的年份是 2008 年,那么 $A=9$, $B=46.76\%$, $C=5.16\%$,最终计算 $hl=9.627$,时间以年为单位,计算结果如表 5 所示:

表 5　YYT 科研偏好部分标签及其权重

科研偏好标签	偏好权重 ω_i	衰减权重 θ_i	综合权重
片上网络	0.007 5	0.604 1	0.305 7
异步传输协议	0.007 8	0.523 0	0.265 4
衬底驱动	0.006 2	0.523 0	0.264 6
低功耗	0.009 1	0.392 1	0.200 6
CMOS	0.020 8	0.339 6	0.180 2
延时电路	0.005 2	0.336 0	0.172 4
温度响应	0.013 0	0.273 6	0.143

3　科研关系的标签用科研合作者的姓名来表示。笔者从获取的科研人员成果信息中,整合其中的所有作者,最终获得产生科研关系的人员有 1085 个,根据公式(3)－(7)分别计算每位作者的贡献度 R_u 和科研关系强度 R_{ui},计算结果见表 6。那么,该科研人员的科研关系模型可表示为:

$$Relation=\{<YYT,261.5498,2.98>,<ZZM,105.0262,2.88>,<YJJ,101.7000,1.96>\cdots\}$$

表 6　YYT 科研关系部分标签及其权重

Re*Tag*	R_{ui}	R_u	Re*Tag*	R_{ui}	R_u
YYT	261.549 8	0.073 2	DRX	48.981 0	0.006 41
ZZM	105.026 2	0.031 6	CCC	39.681 2	0.506 41
YJJ	101.700 0	0.032 8	DG	33.961 9	0.006 41
YZ	83.066 7	0.010 7	GXG	33.609 5	0.004 90
LYJ	59.641 3	0.009 8	CJG	30.095 2	0.004 09
WJY	56.378 6	0.008 7	LLX	24.652 4	0.003 91

基于科研关系模型,构建科研人员关系网络,R_{ui} 作为边权,R_u 作为点权,将标签信息和权重信息导入到 Pajek 中,生成 YYT 的科研关系网络图(见图 3)。点击任意节点,即可查看该科研人员的画像。

3.4　科研人员画像可视化

本文采用可视化工具 Tagul,将所得到的标签导入到 Tagul 中,依据标签的权重设置标签的大小。在目前成熟的可视化工具中以及现有的关于标签云的研究

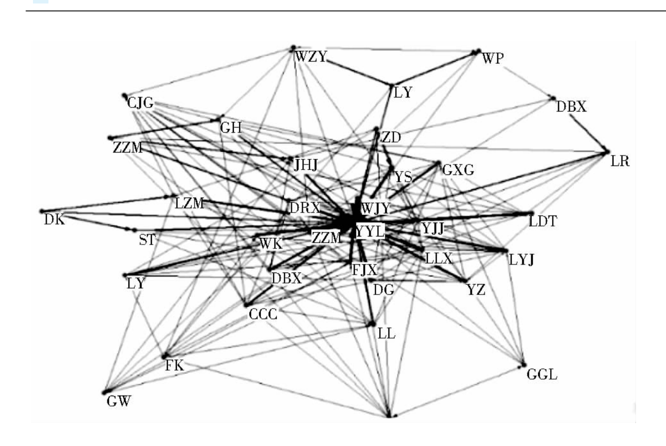


图 3　YYT 科研关系网络图

中,标签的设置多依据个人的主观意愿,还未出现统一的设置标准,本文为了使科研人员画像展示更加美观,将标签大小设置在 10 以内,以便更好地展示科研人员信息。基础特征在了解科研人员信息时是比较重要的,也是变化最小的,设置基础特征的标签大小为 10。科研偏好标签和科研关系标签的大小,用公式 $Size = 10 \times \frac{\text{某个标签权重}}{\text{所有标签权重之和}}$来计算,这样所有的标签大小控制在 10 以内。最后,以该科研人员的照片为背景,制作科研人员画像,见图 4。用同样的方法最终形成 P. Domingos 的科研人员画像,见图 5。在实际应用过程中,可以用不同颜色来区分各类标签,使得信息展示更加一目了然。



图 4　YYT 科研人员画像可视化展示

4　讨论

本文基于多源科技管理数据,提出了科研人员画像的构建方法。该方法从基本属性、科研兴趣以及科研关系方面描述了科研人员信息,并借助标签云的原理,把科研人员信息直观地展示出来,使得科技管理者在科技决策时能够快速掌握科研人员信息,有效提高决策效率。针对科研人员画像构建过程中的一些问题,进行以下讨论:



图 5　P. Domingos 科研人员画像可视化展示

在对科研人员个人简介信息进行命名实体识别时,采用了基于词典的识别方法。但是,目前针对于命名识别的研究有很多。本文的实例验证中采用基于词典的识别方法,词典的完备程度是影响信息抽取效果的重要因素。本文为了使信息抽取结果更加准确,将类似于“西安电子科技大学”·这样的长词添加到词典中,但是在处理多个科研人员信息时,逐个添加词到词库中的方法会浪费大量的时间,当处理批量的信息抽取时,可以结合机器学习算法如 CRF 算法进行抽取[27-29],并且 CRF 算法在命名识别时已经取得了良好的效果。

针对实例验证中的关键词中英对照问题,现有的研究已经实现了自动化翻译,常用的方法有基于规则的[30]、基于实例的[31]、基于统计的[32]和基于神经网络的[33,34]方法。近年来,基于深度学习的神经机器翻译方法[35]获得迅速发展,成为学术界和工业界新的主流方法。但是,不同的机器翻译方法的准确度还有待提高。在本文的实例中,为了使构建的科研人员画像更加准确,采用了人工修正的方法,今后随着技术的不断完善,当处理大量数据时,可以采用自动化翻译实现大规模的中英关键词的对照。

5　结语

针对目前研究中科研人员信息描述角度单一和展示不直观的问题,本文首先提出了科研人员画像的概

念,认为其是一个标签化的形式化描述科研人员的用户模型,然后从科研人员的基础属性、科研兴趣以及科研关系维度构建了科研人员画像,最后以可视化的方式展现出来,并以国内外科研人员为例验证了该方法的可行性。该方法既全面描述了科研人员信息,也使得信息的展示更加直观,在科研人员的评价、评审专家推荐以及团队组建中有重要作用。本文还存在以下不足:不同的科研人员侧重的信息也不同,基础性的研究人员侧重考虑其成果的水平和质量以及在业界的影响,而应用型研究的科研人员侧重考虑其成果的转化价值,探究不同类型的科研人员画像是本文研究需要拓展的方面。

参考文献:

[1] 刘学,刘颖,康运廷,等.国立科研机构科研人员评价框架[J].科研管理,2015,36(S1):224－226.

[2] 王飞跃. 知识产生方式和科技决策支撑的重大变革——面向大数据和开源信息的科技态势解析与决策服务[J]. 中国科学院院刊, 2012, 27(5):527－537.

[3] 化柏林.科技信息大数据在情报研究服务中的应用[J].图书情报工作,2017,61(16):150－156.

[4] 宋培彦, 陈白雪, 贤信. 科技专家信息语义模型构建及实证研究[J]. 情报理论与实践, 2017, 40(9):119－124.

[5] EuroCRIS[EB/OL].[2016－02－15]. http://www. eurocris. org/.

[6] MOREIRA C, WICHERT A. Finding academic experts on a multi-sensor approach using Shannon's entropy[J]. Expert systems with applications, 2013, 40(14):5740－5754.

[7] YANG L, HU Z, LONG J. Service of searching and ranking in a semantic-based expert information system[C]//Proceedings of the IEEE Asia-Pacific services computing conference. Los Angeles: IEEE Computer Society, 2010:609－614.

[8] DUONG T H, UDDIN M N, NGUYEN C D. Personalized semantic search using ODP: astudy case in academic domain[C]//Proceedings of theinternational conference on computational science and itsapplications. Vietnam: Springer Berlin Heidelberg, 2013:607－619.

[9] 李纲,叶光辉.多源专家特征信息融合研究[J].现代图书情报技术,2014(4):27－33.

[10] 王曰芬, 王雪芬, 杨小晓. 基于社会网络的科技咨询专家库的构建方案与流程设计[J]. 情报学报, 2012, 31(2):116－125.

[11] 陆伟, 韩曙光. 组织专家的检索系统设计与实现[J]. 情报学报, 2008, 27(5):657－663.

[12] TANG J, YAO L, ZHANG D, et al. A combination approach to Web user profiling[J]. ACMtransactions on knowledge discovery from data, 2010, 5(1):1－44.

[13] 化柏林, 李广建. 大数据环境下多源信息融合的理论与应用探讨[J]. 图书情报工作, 2015,59(16):5－10.

[14] COOPER A. 交互设计之路[M]. DING C,译.北京:电子工业出版社,2006.

[15] QIU X, ZHANG Q, HUANG X. FudanNLP: a toolkit for Chinese natural language processing[C]// Proceedings of the meeting of the Association for Computational Linguistics: system demonstrations. Sofia: the Association for Computational Linguistics,2013: 49－54.

[16] ZHOU L, ZHANG D. NLPIR: atheoretical framework for applying natural language processing to information retrieval[J]. Journal of the American Society for Information Science & Technology, 2003, 54(2):115－123.

[17] 刘挺,车万翔,李正华.语言技术平台[J].中文信息学报,2011, 25(6):53－62.

[18] 宋文强. 科技文献作者重名消歧与实体链接[D].哈尔滨:哈尔滨工业大学, 2012.

[19] 章顺瑞,游宏梁.基于层次聚类算法的中文人名消歧[J].现代图书情报技术,2010(11):64－68.

[20] 朱云霞.中文文献题录数据作者重名消解问题研究[J].图书情报工作,2014,58(23):143－148,142.

[21] CHENG Y, QIU G, BU J, et al. Model bloggers' interests based on forgetting mechanism[C]//Proceedings of theinternational conference on World Wide Web. New York: ACM, 2008:1129－1130.

[22] 樊玉敬. 合著论文作者的名誉分配[J]. 情报杂志, 1997(1): 37－38.

[23] 赵树庆, 刘永胜. 20 种科技期刊 2010 年论文发表时滞调查[J]. 编辑学报, 2011, 23(6):491－493.

[24] WordArt. com - Word Cloud Art Creator[EB/OL].[2018－01－30]. https://wordart. com/.

[25] Tagxedo - Word Cloud with Styles[EB/OL].[2018－01－30]. http://www. tagxedo. com/.

[26] 陈木佩.学术关键词半衰期测度和老化研究[J].科技创业月刊,2010,23(8):156－157.

[27] 徐建忠,朱俊,赵瑞,等.基于 CRF 算法的航天命名实体识别[J].电子设计工程,2017,25(20):42－46.

[28] 谢志宁. 中文命名实体识别算法研究[D].杭州:浙江大学, 2017.

[29] 曾冠明. 基于条件随机场的中文命名实体识别研究[D].北京:北京邮电大学,2009.

[30] 刘颖,姜巍.基于翻译规则的统计机器翻译[J].计算机科学, 2013,40(2):214－217.

[31] 刘占一,李生,刘挺,等.利用统计搭配模型改进基于实例的机器翻译[J].软件学报,2012,23(6):1472－1485.

[32] 赵静.基于统计的汉英机器翻译技术的研究[J].电子设计工程,2016,24(21):69－71,75.

[33] 丁亮,何彦青.融合领域知识与深度学习的机器翻译领域自适应研究[J].情报科学,2017,35(10):125－132.

[34] 李婧萱. 基于深度神经网络的统计机器翻译模型研究[D].哈尔滨:哈尔滨工业大学,2016.

[35] 杨南. 基于神经网络学习的统计机器翻译研究[D].合肥:中

国科学技术大学,2014.

作者贡献说明:
范晓玉:提出论文初步研究思路,起草论文;
窦永香:指导论文总体研究思路,多次修改论文;
赵捧未:指导论文思路;
周潇:修改论文。

Study for the Construction Method of Scientist Profile with Multi-Source Data Fusion

Fan Xiaoyu Dou Yongxiang Zhao Pengwei Zhou Xiao
School of Economics and Management, Xidian University, Xi'an 710071

Abstract: [**Purpose/significance**] In the age of big data, people need to be digitized, and researchers need to digitize them. The establishment of scientists profile is of great importance for scientific research managers to comprehensively understand the information of researchers and objectively evaluate their research level. It can be used as the basis for analyzing the research behavior or recommendation of experts, and improving the efficiency of scientific research management. [**Method/process**] First of all, the concept of scientists profile is proposed, which is a collection of labels describing the information of scientific researchers. Secondly, based on the data of multiple heterogeneous data sources, such as personal homepage, knowledge network and fund network, this paper proposes a method for the construction of scientists profile with multi-source data. This method formally describes the information of scientific researchers from the three aspects of the basic attribute, scientific research preference and scientific research relationship, and extracts the labels of each dimension to vividly display the profile in a visual way. Finally, the feasibility of this method is illustrated by taking two researchers at home and abroad as examples. [**Result/conclusion**] The construction of the scientists profile is suitable for researchers at home and abroad, which can fully describe the information of researchers and show them visually.

Keywords: scientist profile multi-source data user model